Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/US05/000517

International filing date: 07 January 2005 (07.01.2005)

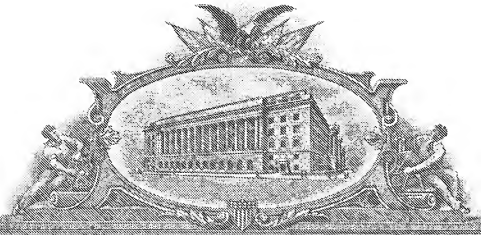
Document type: Certified copy of priority document

Document details: Country/Office: US
Number: 60/535,111
Filing date: 08 January 2004 (08.01.2004)

Date of receipt at the International Bureau: 14 March 2005 (14.03.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)

World Intellectual Property Organization (WIPO) - Geneva, Switzerland
Organisation Mondiale de la Propriété Intellectuelle (OMPI) - Genève, Suisse



1291948

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

March 03, 2005

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE.

APPLICATION NUMBER: *60/535,111*
FILING DATE: *January 08, 2004*
RELATED PCT APPLICATION NUMBER: *PCT/US05/00517*

Certified by

Jon W. Dudas

Under Secretary of Commerce
for Intellectual Property
and Director of the United States
Patent and Trademark Office

PTO/SB/16 (08-03)
Approved for use through 07/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

PROVISIONAL APPLICATION FOR PATENT COVER SHEET

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53(c).

Express Mail Label No. EL 981814135 US

INVENTOR(S)

Given Name (first and middle [if any])	Family Name or Surname	Residence (City and either State or Foreign Country)
Alicia	Bertone	Columbus, OH

Additional inventors are being named on the 1 *separately numbered sheets attached hereto*

TITLE OF THE INVENTION (500 characters max)

METHODS OF USING PUBLIC DATABASES TO CREATE GENE EXPRESSION MICROARRAYS, AND MICROARRAYS CREATED THEREBY

Direct all correspondence to: **CORRESPONDENCE ADDRESS**

☑ Customer Number: 27874

OR

☐ Firm or Individual Name

Address

Address

City | State | Zip

Country | Telephone | Fax

ENCLOSED APPLICATION PARTS *(check all that apply)*

☑ Specification *Number of Pages* 1386

☐ CD(s), Number

☑ Drawing(s) *Number of Sheets* 1

☑ Other (specify) App. A, Tables 5-10

☐ Application Date Sheet. See 37 CFR 1.76

METHOD OF PAYMENT OF FILING FEES FOR THIS PROVISIONAL APPLICATION FOR PATENT

☐ Applicant claims small entity status. See 37 CFR 1.27.

☑ A check or money order is enclosed to cover the filing fees.

☐ The Director is herby authorized to charge filing fees or credit any overpayment to Deposit Account Number:

☐ Payment by credit card. Form PTO-2038 is attached.

FILING FEE Amount ($)

$80.00

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

☑ No.

☐ Yes, the name of the U.S. Government agency and the Government contract number are:

Respectfully submitted,

SIGNATURE Sean Myers-Payne

TYPED or PRINTED NAME Sean C. Myers-Payne

Date January 8, 2004

REGISTRATION NO. 42,920 *(if appropriate)*

Docket Number: 18525/04099

TELEPHONE 614-621-7754

USE ONLY FOR FILING A PROVISIONAL APPLICATION FOR PATENT

PROVISIONAL APPLICATION COVER SHEET
Additional Page

PTO/SB/16 (08-03)
Approved for use through 07/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

Docket Number 18525/04099

INVENTOR(S)/APPLICANT(S)

Given Name (first and middle [if any])	Family or Surname	Residence (City and either State or Foreign Country)
Weisong	Gu	Dublin, OH

Number 1 of 1

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

UNITED STATES PROVISIONAL PATENT APPLICATION

FOR

METHODS OF USING PUBLIC DATABASES TO CREATE GENE EXPRESSION MICROARRAYS, AND MICROARRAYS CREATED THEREBY

BY

ALICIA BERTONE

WEISONG GU

METHODS OF USING PUBLIC DATABASES TO CREATE GENE EXPRESSION MICROARRAYS, AND MICROARRAYS CREATED THEREBY

DESCRIPTION OF THE INVENTION

Field of the Invention

[001] The present invention is directed to methods of preparing biological databases, and databases prepared according to those methods. In some embodiments, the methods can be performed entirely using computer resources, relying solely on publicly available biological sequence information. The methods of the invention can be used to generate species-specific nucleic acid microarrays.

Background of the Invention

[002] DNA microarrays are small, solid supports containing thousands of different gene sequences that are immobilized or attached at fixed locations. (Ekins R and Chu FW, "Microarrays: their origins and applications," *Trends Biotechnol* 17:217-218 (1999); Lobenhofer EK, Bushel PR, Afshari CA, and Hamadeh HK, "Progress in the Application of DNA Microarrays," *Environ Health Perspect* 109(9):881-891 (2001).) This technology has revolutionized the basic approach to research since its invention. Unlike the traditional methods in molecular biology for one gene in one experiment, hundreds to thousands of genes can be analyzed simultaneously under identical conditions to various biological models, including disease, therapy, or experimental manipulation. Microarrays provide unprecedented opportunities for both qualitative and quantitative analysis in gene expression, gene identification and gene alteration detection, such as polymorphisms. (Galamb O, Molnar B, and Tulassay Z, "DNA chips for gene expression analysis and their

application in diagnostics," *Orv Hetil* 144:21-27 (2003)). The use of larger scale expression profiling permits the classification of genes by biological function, the contribution of patients' disease patterns directly to research, as well as the discovery of genes of unknown function by association with disease. The expression profiles can be diagnostic, prognostic, as well as disease monitoring. (Bubendorf L, "High-throughput microarray technologies: from genomics to clinics," *Eur. Urol* 40:231-238 (2001); Crowther DJ, "Applications of microarrays in the pharmaceutical industry," *Curr. Opin. Pharmacol* 2:551-554 (2002).)

[003] Mammalian commercial DNA microarrays currently exist for human, mouse, and rat, but not for the horse, and few other domestic animals.

[004] There are currently two dominant DNA microarray technologies: spotted microarrays on glass slides, which were first developed at Stanford University (Schena M, Shalon D, Davis RW, and Brown PO, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science* 270:467-470 (1995)), and *in situ* synthesized oligonucleotide microarrays produced by Affymetrix Inc. Spotted microarrays contain probes that are complementary DNA (cDNA), polymerase chain reaction products or oligonucleotides. Probes are physically deposited on a chemically modified glass slide. Two purified mRNA samples are separately reverse transcribed using two different fluoroprobes and the resulting dye-labeled cDNA populations are used to hybridize on the array under competitive conditions. After hybridization, the array is analyzed with a two-channel fluorescence scanner and the ratio of the two fluorophores can be determined which is later used to reflect the gene expression level of target genes. (Burgess JK,

"Gene expression studies using microarrays," *Clin Exp Pharmacol Physiol* 28(4):321-328 (2001).) One of the major advantages of cDNA spotted microarrays is that the genetic information need not be known before putting it on the array. Yet if the genetic information is available, oligonucleotides can be specifically designed to uniquely hybridize the target gene.

[005] The Affymetrix arrays, on the other hand, are manufactured through a unique process -- a combination of photolithography and combinatorial chemistry -- that results in many of the arrays' capabilities. With a calculated minimum number of synthesis steps, GeneChip technology produces arrays with hundreds of thousands of different probes packed at an extremely high density. This feature enables researchers to obtain high quality, genome-wide data using small sample volumes. Manufacture is scalable because the length of the probes, not their number, determines the number of synthesis steps required. This production process yields arrays with highly reproducible properties, which reduces user set-up time by eliminating the need for individual labs to produce and test their own arrays.

[006] Here, we describe a unique computer-based approach for the data mining and sequence selection for the gene expression microarray from the GenBank database using a series of Java application programs.

SUMMARY OF THE INVENTION

[007] The present invention is advantageous in providing a new method for obtaining a species-specific collection of nucleic acid sequences from publicly available databases. In particular, the present invention provides methods of preparing a species-specific nucleic acid database comprising: selecting from a species-non-specific nucleic acid database species-specific nucleic acids comprising coding sequences; selecting from a species-non-specific nucleic acid database species-specific nucleic acids comprising noncoding sequences; selecting from the coding sequences those sequences that are 3′-complete or 3′-coding biased, wherein 3′-coding biased sequences comprise 5′-partial sequences having desirable characteristics; selecting from the noncoding sequences those sequences that include poly-A tails or are derived from sequences that include poly-A tails; reducing redundancy in selected sequences; comparing sequences comprising unannotated sequences to a collection of sequences comprising annotated coding sequences and selecting those sequences satisfying a threshold of similarity; and collecting all selected sequences. In some embodiments, the species-specific nucleic acid database is an equine-specific nucleic acid database. In some embodiments, the species-non-specific nucleic acid database is GenBank. In some embodiments, the databases according to the present invention can be used in the diagnosis of various animal diseases, including human, equine, or canine diseases. The diagnostic tests in equine and canine can be used to model human diseases.

[008] The present invention also provides arrays comprising a plurality of oligonucleotide probes designed to be complementary to and hybridize under stringent conditions with a gene listed in one of Tables 5, 7, or 9.

[009] The present invention also provides arrays comprising a plurality of oligonucleotides, wherein: a) the oligonucleotides are chosen from the nucleic acid sequences shown in Tables 6, 8, or 10, and wherein the array comprises 10 or more of said oligonucleotides; or b) the oligonucleotides comprise nucleotide probes designed to be complementary to, or hybridize under stringent conditions with, 10 or more nucleic acid sequences shown in Tables 6, 8, or 10. In some embodiments, the oligonucleotides comprise nucleotide probes designed to be complementary to, or hybridize under stringent conditions with, 1000, 2000, or 3000 or more nucleic acid sequences shown in Table 6.

[010] The present invention also provides methods for populating a database of species-specific nucleic acid sequences, comprising querying a database of nucleic acid sequences to identify nucleic acid sequences associated with a subject species; processing the identified sequences to create a first subset containing coding sequences and a second subset containing non-coding sequences; dividing the first subset into a plurality of DNA sequences, if present, and a plurality of mRNA sequences; processing the plurality of DNA sequences to derive a plurality of virtual mRNA sequences; dividing the plurality of mRNA sequences into a plurality of complete and mRNA 3' partial sequences, and a plurality of mRNA 5' partial sequences; processing the plurality of mRNA 5' partial sequences to identify a subset of mRNA 5' partial sequences, each member of the subset satisfying a

threshold level of completeness; identifying members of the second subset containing non-coding sequences that correlate with at least one known coding sequence of at least one species other than the subject species; and combining the plurality of virtual mRNA sequences, the plurality of complete and mRNA 3' partial sequences, the subset of mRNA 5' partial sequences, and the identified correlated sequences to create the database of species-specific nucleic acid sequences. In some embodiments, the step of identifying includes comparing each member of the second subset to each member of a database containing annotated human nucleic acid sequences. In some embodiments, the step of identifying includes comparing each member of the second subset to each member of a database containing annotated human and mouse nucleic acid sequences. The database containing annotated human and mouse nucleic acid sequences can be derived from the database of nucleic acid sequences. In some embodiments, the method further comprises eliminating duplicates within the database of species-specific nucleic acid sequences. In some embodiments, the method further comprises populating the database of species-specific nucleic acid sequences with selected species-specific virus definitions. In some embodiments, the method further comprises verifying that each of the identified correlated sequences is represented in sense format.

[011] With regard to the methods of the present invention, it is contemplated that fewer or more steps may be included, or that steps may be combined, or that steps may be rearranged.

[012] Additional aspects and advantages of the invention will be set forth in part in the description that follows, and in part will be obvious from the description, or

may be learned by practice of the invention. The objects and advantages of the invention will be realized and attained by means of the elements and combinations particularly pointed out in the appended claims.

[013] It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory only and are not restrictive of the invention, as claimed.

[014] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate one (several) embodiment(s) of the invention and together with the description, serve to explain the principles of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[015] Figure 1 is a schematic flow chart of the overall design of 3′-biased equine annotated gene and EST sequence selection.

DESCRIPTION OF THE EMBODIMENTS

[016] Reference will now be made in detail to specific embodiments of the invention, examples of which are illustrated in the accompanying drawings. Wherever possible, the same reference numbers will be used throughout the drawings to refer to the same or like parts.

[017] The present invention is generally directed to methods for preparing biological databases, and databases prepared according to those methods. The inventive methods can be practiced using readily available hardware and publicly available software. The databases can comprise nucleic acids, including DNA and/or RNA, or polypeptides.

[018] In one embodiment, the invention comprises methods for curating, pruning, and annotating publicly available gene sequences by computer to create high quality nucleic acid sequence data. The data obtained by the present methods can be assembled into a database, which can be used for any purpose, including use in a gene expression microarray.

[019] The methods of the invention take advantage of information available in public databases, including but not limited to, GenBank. As will be readily apparent from this disclosure, other databases can also be used, provided the desired information is available.

[020] The methods of the invention can accommodate selection of any desired characteristics of the nucleic acid sequences. For example, the invention can be used to select all species-specific sequences, such as all equine (Equus caballus), bovine (Bos taurus), ovine (Ovis aries), porcine (Sus scrofa), caprine (Capra hircus), canine (Canis familiaris), feline (Felis catus), avian (domestic chicken, Gallus gallus), or any other desired species. Within any given species, selection can be all inclusive or be made based on tissue, or disease, or pathogen, or any other desired characteristic.

[021] The invention will now be described with reference to a particular embodiment. It should be recognized that the invention comprises other embodiments, and that those of ordinary skill in the art will recognize what those embodiments are. Also, the embodiments described herein comprise several steps or components. It is contemplated that these steps may be rearranged, as desired, to achieve the desired result. Additional or fewer steps can also be performed. The

numbering scheme below is simply for clarification in this description and is not intended to define the order of the steps.

[022] Additionally, while the following steps are designed for selecting mRNA sequences, other selections could be made during any step, depending on the desired result. Finally, the following steps selected for 3′-biased mRNA sequences, but other selection forces may be applied, including for example, selecting for *all* mRNA sequences, selecting for DNA sequences, selecting for complete sequences, etc. The choices will be understood by those of skill in the art upon reading this disclosure.

[023] 1. Obtaining a Species-Specific Selection of Nucleic Acid Sequences

[024] In one embodiment of the invention, a species-specific collection of nucleic acid sequences is prepared. In a first step, a database, such as the publicly available GenBank database, is queried using a species-specific request. (The method of this invention can be applied to any database, including proprietary databases.) For example, to obtain all equine sequences, the database is queried for "Equus caballus," for bovine, "Bos taurus," for ovine, "Ovis aires," for porcine, "Sus scrofa," for caprine, "Capra hircus," for canine, "Canis familiaris," or for feline, "Felis catus."

[025] It should be recognized that public databases may differ in the information that may be entered for any given field. For example, instead of simply "Equus caballus," an entry may say "Equus caballus (horse)," or other similar entry. Thus, if desired, care may be taken to use inclusive language in the query to avoid omitting desired entries. Similarly, it should be recognized that entries may refer to a

species as a host, such as "Equine lymphoma." If desired, care can be taken to use exclusive language to avoid including such entries.

[026] In some embodiments, the program GetEquine is used to select equine sequences from the initial collection of sequences.

[027] 2. Separating Coding Sequences (CDS) from Non-Coding Sequences (NonCDS)

[028] The Coding Sequences (CDS) and Non-Coding Sequences (NonCDS) can then be separated. In some embodiments, they are separated by the program GetCDS. NonCDS can undergo further analysis, as described herein below in step 11. Within the CDS selection, some sequences may comprise DNA and others mRNA.

[029] 3. Separation of DNA CDS from mRNA CDS

[030] By the program CheckMRNA, one can separate mRNA sequences from DNA sequences. Sequences identified as "mRNA" are treated further below under step number 7. DNA CDS may further comprise complete and partial sequences.

[031] 4. Selection of 3′ Complete DNA Sequences

[032] "Complete 3′" DNA coding sequences contain stop codons at the three-prime ends, and thus can be full-length or partial sequences anchored at their three-prime ends. Other sequences are 5′ partial DNA sequences. The DNA CDS from step 3 above can be further selected for "3′ complete" sequences, to remove 5′ partial sequences from the collection. Of course, if desired, partial DNA sequences can be retained and later analyzed and annotated.

[033] 5. Removing Duplicate Sequences

[034] Because there is a possibility that multiple entries exist for the same sequence, steps may be taken to remove duplicates. In the case of GenBank sequences, the selected DNA sequences from step 4 can be converted to a uniform format, such as Fasta format by using the FastaG program, then submitted to an overlap-detecting algorithm, such as the ClusterG program. Any level of scrutiny can be applied in identifying "duplicates." For example, sequences that are greater than 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, 50%, or even lower percent, identical can be deemed duplicates and removed. Obviously, a higher level allows for a larger number of similar sequences to be retained, whereas a lower level will have the opposite effect. The desired level can be unique to any situation, and will be determined by the scientist or practitioner using the system, depending on their needs.

[035] 6. Identifying "Buried" mRNA Sequences

[036] The non-duplicate DNA CDS can further be examined for the presence of mRNA information. When available, the mRNA information can be collected and further analyzed as described below step number 10.

[037] 7. Selection of 3´ Complete mRNA Sequences

[038] Like the DNA described above, "3´ Complete" mRNA coding sequences contain stop codons and/or a polyA-tail at the three-prime ends, and thus can be full-length or partial sequences anchored at their three-prime ends. Other sequences are 5´ partial mRNA sequences. The mRNA CDS from step 3 above can be further selected for "3´ complete" sequences, to remove 5´ partial sequences

from the collection. Unlike with partial DNA sequences, however, partial mRNA sequences are retained for further processing as described in step 9, below. (Partial DNA sequences can be retained if desired.)

[039] 8. Removing Duplicate Sequences

[040] Because there is a possibility that multiple entries exist for the same sequence, steps may be taken to remove duplicates. In the case of GenBank sequences, the selected complete 3′ mRNA sequences from step 7 above can be converted to a uniform format, such as Fasta format by using the FastaG program, then submitted to an overlap-detecting algorithm, such as the ClusterG program. Any level of scrutiny can be applied in identifying "duplicates." For example, sequences that are greater than 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, 50%, or even lower percent, identical can be deemed duplicates and removed. Obviously, a higher level allows for a larger number of similar sequences to be retained, whereas a lower level will have the opposite effect. The desired level can be unique to any situation, and will be determined by the scientist or practitioner using the system, depending on their needs. Sequences selected are further treated in step 10, below.

[041] 9. Annotating Partial mRNA Sequences

[042] Because 5′ partial mRNA from step 7 above may include regions close to the 3′ end, and thus be suitable for use in a microarray, further analysis of these sequences can be performed.

[043] First, the 5′ partial mRNA from step 7 are compared to a combined coding sequence database, such as human + mouse, which can be obtained by

querying GenBank for "homo cds" and combining those results with "mus cds." The coding sequence database can include any sequences, but highly evolved and annotated databases are desirable as the comparative database. The comparison can be achieved using a sequence comparison program such as "BlastN." The program compares sequences and identifies those that are similar or identical. As with similar programs, the stringency of the comparison can be varied, so as to be more or less selective. Thus, a Blast "score" can be greater than 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, or higher, depending on the desire for identifying similar or identical sequences. Another measurement that can be used is the "E" value, which can be less than 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}, 10^{-10}, or even less, again depending on the desire for identifying similar or identical sequences.

[044] Sequences can then be further selected for their closeness to the 3′ end. "Closeness" is a subjective determination, but can be arbitrarily set at any number of bp, such as less than 1000 bp, 900, 800, 700, 600, 500, 400, 300, 200, 100, or fewer bp, from the 3′ end.

[045] 10. Combining and Processing Selected Species Sequences

[046] "Buried" mRNA sequences from step 6, 3′ complete mRNAs from step 8, and selected 5′ partial mRNAs from step 9 are combined, and further processed for duplicates. Again, the sequences can be converted to a uniform format, such as Fasta format by using the FastaG program, then submitted to an overlap-detecting algorithm, such as the ClusterG program. Any level of scrutiny can be applied in identifying "duplicates." For example, sequences that are greater than 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, 50%, or even lower

percent, identical can be deemed duplicates and removed. Obviously, a higher level allows for a larger number of similar sequences to be retained, whereas a lower level will have the opposite effect. The desired level can be unique to any situation, and will be determined by the scientist or practitioner using the system, depending on their needs. The selected sequences are further processed as described in step 15, below.

[047] 11. Selection of Poly-A ESTs from Non-CDS

[048] Because Non-CDS may still include useful sequences, the Non-CDS from step 2 above can be further processed. The Non-CDS are further selected for those that are identified as including a poly-A tail. This can be performed by running a program such as GetPolyAEST on the NonCDS collection of sequences. The sequence information from these ESTs may contain the polyA tail if the sequencing process reaches to the 3′ end. However, if the sequencing is initiated at the 5′ end and stops in the middle, the obtained sequence information may not include the polyA tail, although it may be very close to the 3′ end. Therefore, ESTs claiming "PolyA=No" may not necessarily mean that they are not at or close to the 3′ end. Based on this, we first selected the ESTs which claim both "PolyA=Yes" and "PolyA=No" so that a maximal pool of candidate 3′ ESTs could be constructed.

[049] 12. Selection of High Quality ESTs

[050] The poly-A-containing ESTs from step 11 above are further processed to select high-quality, vector-trimmed regions. For example, in Genbank there is a feature that states the regions that are of high phred quality with the start and stop positions. All sequences were trimmed to only include these high quality regions

based on the start and stop positions. This enhances the confidence that the sequencing was completed accurately.

[051] 13. Removing Duplicate Sequences

[052] Again, because there is a possibility that multiple entries exist for the same sequence, steps can be taken to remove duplicates. Removal of duplicates can achieve any practical purpose, including, for example, maximizing the space limitations of a microarray, or simply reducing the costs of producing the microarray. In the case of GenBank sequences, the selected poly-A ESTs from step 12 above can be converted to a uniform format, such as Fasta format by using the FastaG program, then submitted to an overlap-detecting algorithm, such as the ClusterG program. Any level of scrutiny can be applied in identifying "duplicates." For example, sequences that are greater than 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, 50%, or even lower percent, identical can be deemed duplicates and removed. Obviously, a higher level allows for a larger number of similar sequences to be retained, whereas a lower level will have the opposite effect. The desired level can be unique to any situation, and will be determined by the scientist or practitioner using the system, depending on their needs.

[053] 14. Annotating poly-A EST Sequences

[054] The polyA ESTs can be compared to a coding sequence database, such as a combined human + mouse coding sequence database, which can be obtained by querying GenBank for "mus cds" and combining those results with "homo cds." The comparison can be achieved using a sequence comparison

program such as "BlastN." The program compares sequences and identifies those that are similar or identical. As with similar programs, the stringency of the comparison can be varied, so as to be more or less selective. Thus, a Blast "score" can be greater than 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, or higher, depending on the desire for identifying similar or identical sequences. Another measurement that can be used is the "E" value, which can be less than 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}, 10^{-10}, or even less, again depending on the desire for identifying similar or identical sequences.

[055] Sequences with lower blast score values than the set threshold can be subjected to further analysis. For example, If there is a polyA tail or polyA signals (AATAAA or ATTAAA) present at the 3' end, although with low or 0 blast score, that sequence can also be selected and considered as sense.

[056] Sequences with above-threshold blast scores can then be further selected for their closeness to the 3´ end. "Closeness" is a subjective determination, but can be arbitrarily set at any number of bp, such as less than 1000 bp, 900, 800, 700, 600, 500, 400, 300, 200, 100, or fewer bp, from the 3´ end.

[057] Still further, the sense or anti-sense orientation of the sequence can be determined, for example, through use of the BlastN program, which shows the direction of the match. Those sequences deemed to be in anti-sense orientation can be converted to sense sequences by, for example, programs that reverse complement the sequence.

[058] The selected sense-oriented 3′-biased ESTs and converted anti-sense 3′-biased ESTs can be combined together and further processed as described below in step 15.

[059] 15. Combining Sequences and Removing Duplicates

[060] The selected sequences from step 10 are combined with those selected from step 14. To reduce the existence of duplicates, further processing can be performed. The selected sequences can be converted to a uniform format, such as Fasta format by using the FastaG program, then submitted to an overlap-detecting algorithm, such as the ClusterG program. Any level of scrutiny can be applied in identifying "duplicates." For example, sequences that are greater than 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, 50%, or even lower percent, identical can be deemed duplicates and removed. Obviously, a higher level allows for a larger number of similar sequences to be retained, whereas a lower level will have the opposite effect. The desired level can be unique to any situation, and will be determined by the scientist or practitioner using the system, depending on their needs.

[061] The collection of data created in the steps above can be used for any applicable purpose. Those of skill in the art will recognize uses for such information. The nucleic acid sequences can be used as they are or transformed for any desired use. For example, the sequences can be translated into polypeptide sequences, which can be used for any desired purpose, or probes can be derived from the nucleic acid sequences selected.

[062] Polynucleotide Probes

[063] Probes can be genomic DNA or cDNA or mRNA, or any RNA-like or DNA-like material, such as peptide nucleic acids, branched DNAs and the like. Probes can be sense or antisense polynucleotide probes. Where target

polynucleotides are double stranded, the probes may be either sense or antisense strands. Where the target polynucleotides are single stranded, the nucleotide probes are complementary single strands.

[064] Probes can be prepared by a variety of synthetic or enzymatic schemes, examples of which are well known in the art. Probes can be synthesized, in whole or in part, using chemical methods, examples of which are well known in the art (Caruthers et al. (1980) Nucleic Acids Res. Symp. Ser. 215-233). Alternatively, the probes can be generated, in whole or in part, enzymatically.

[065] Nucleotide analogs can be incorporated into polynucleotide probes by methods well known in the art. The incorporated nucleotide analogues should serve to base-pair with target polynucleotide sequences. For example, certain guanine nucleotides can be substituted with hypoxanthine, which base-pairs with cytosine residues. However, these base pairs may be less stable than those between guanine and cytosine. Alternatively, adenine nucleotides can be substituted with 2,6-diaminopurine, which can form stronger base pairs than those between adenine and thymidine. Additionally, polynucleotide probes can include nucleotides that have been derivatized chemically or enzymatically. Typical chemical modifications include derivatization with acyl, alkyl, aryl, or amino groups.

[066] The probes can be labeled with one or more labeling moieties to allow for detection of hybridized probe/target polynucleotide complexes. The labeling moieties can include compositions that can be detected by spectroscopic, photochemical, biochemical, bioelectronic, immunochemical, electrical, optical, and/or chemical means. The labeling moieties include, for example, radioisotopes,

such as ^{32}P, ^{33}P, or ^{35}S, chemiluminescent compounds, labeled binding proteins, heavy metal atoms, spectroscopic markers, such as fluorescent markers and dyes, magnetic labels, linked enzymes, mass spectrometry tags, spin labels, electron transfer donors and acceptors, and the like.

[067] Probes can be immobilized on a substrate, examples of which include but are not limited to, rigid and/or semi-rigid supports including membranes, filters, chips, slides, wafers, fibers, magnetic or nonmagnetic beads, gels, tubing, plates, polymers, microparticles, and capillaries. Substrates can have a variety of surface forms, such as wells, trenches, pins, channels and pores, to which the probes are bound. The substrates can be optically transparent.

[068] Hybridization complexes

[069] Hybridization causes a probe and a complementary target to form a stable duplex. In the case of polynucleotide probes and targets, this occurs through base pairing. Hybridization methods are well known to those skilled in the art (See, e.g., Ausubel (1997; Short Protocols in Molecular Biology, John Wiley & Sons, New York N.Y., units 2.8-2.11, 3.18-3.19 and 4-6-4.9). Conditions can be selected for hybridization where exactly complementary target and polynucleotide probe can hybridize, i.e., each base pair must interact with its complementary base pair. Alternatively, conditions can be selected where target and polynucleotide probes have mismatches but are still able to hybridize. Suitable conditions can be selected, for example, by varying the concentrations of salt in the prehybridization, hybridization, and wash solutions, or by varying the hybridization and wash

temperatures. With some membranes, the temperature can be decreased by adding formamide to the prehybridization and hybridization solutions.

[070] Hybridization conditions are based on the melting temperature (T_m) of the nucleic acid binding complex or probe, as described in Berger and Kimmel (1987) Guide to Molecular Cloning Techniques, Methods in Enzymology, vol 152, Academic Press. The term "stringent conditions," as used herein, is the "stringency" which occurs within a range from about T_m-5 (5° below the melting temperature of the probe) to about 20°C below T_m. As used herein, "highly stringent" conditions employ at least 0.2 x SSC buffer and at least 65°C. As recognized in the art, stringency conditions can be attained by varying a number of factors, including for example, the length and nature, i.e., DNA or RNA, of the probe; the length and nature of the target sequence; and the concentration of the salts and other components, such as formamide, dextran sulfate, and polyethylene glycol, of the hybridization solution. All of these factors can be varied to generate conditions of stringency which are equivalent to the conditions listed above.

[071] Hybridization can be performed at low stringency with buffers, such as 6xSSPE with 0.005% Triton X-100 at 37°C, which permits hybridization between target and polynucleotide probes that contain some mismatches to form target polynucleotide/probe complexes. Subsequent washes can be performed at higher stringency with buffers, such as 0.5xSSPE with 0.005% Triton X-100 at 50°C, to retain hybridization of only those target/probe complexes that contain exactly complementary sequences. Alternatively, hybridization can be performed with buffers, such as 5xSSC/0.2% SDS at 60°C and washes are performed in

2xSSC/0.2% SDS and then in 0.1xSSC. Background signals can be reduced by the use of detergent, such as sodium dodecyl sulfate, Sarcosyl, or Triton X-100, or a blocking agent, such as salmon sperm DNA.

[072] Other procedures for the use of microarrays are available in the art, and are provided, for example, by Affymetrix. In this regard, reference is made to the Affymetrix GeneChip® Expression Analysis Technical Manual, the entire disclosure of which is incorporated herein by reference.

[073] Microarray Construction

[074] The nucleic acid sequences can be used in the construction of microarrays. Methods for construction of microarrays, and the use of such microarrays, are known in the art, examples of which can be found in U.S. Patent Nos. 5,445,934, 5,744,305, 5,700,637, and 5,945,334, the entire disclosure of each of which is hereby incorporated by reference. Microarrays can be arrays of nucleic acid probes, arrays of peptide or oligopeptide probes, or arrays of chimeric probes -- peptide nucleic acid (PNA) probes. Those of skill in the art will recognize the uses of the collected information.

[075] One particular example, the *in situ* synthesized oligonucleotide Affymetrix GeneChip system, is widely used in many research applications with rigorous quality control standards. (Rouse R. and Hardiman G., "Microarray technology – an intellectual property retrospective," *Pharmacogenomics* 5:623-632 (2003).). Currently the Affymetrix GeneChip uses eleven 25-oligomer probe pair sets containing both a perfect match and a single nucleotide mismatch for each gene sequence to be identified on the array. Using a light-directed chemical

synthesis process (photolithography technology), highly dense glass oligo probe array sets (>1,000,000 25-oligomer probes) can be constructed in a ~ 3 x 3-cm plastic cartridge that serves as the hybridization chamber. The ribonucleic acid to be hybridized is isolated, amplified, fragmented, labeled with a fluorescent reporter group, and stained with fluorescent dye after incubation. Light is emitted from the fluorescent reporter group only when it is bound to the probe. The intensity of the light emitted from the perfect match oligoprobe, as compared to the single base pair mismatched oligoprobe, is detected in a scanner, which in turn is analyzed by bioinformatics software (http://www.affymetrix.com). The GeneChip system provides a standard platform for array fabrication and data analysis, which permits data comparisons among different experiments and laboratories.

[076] Microarrays according to the invention can be used for a variety of purposes, including but not limited to, screening for diseases. For example, sequences can be selected to represent genes of special interest to musculoskeletal systems and neurologic systems as well as the inflammatory and immunologic cascades. The specific intent is to identify gene expression signatures for the presence and degree of diseases, including but not limited to, musculoskeletal diseases such as osteoarthritis, osteochondrosis dessicans, bone disorders of density or healing, and, for example, neurologic diseases, including but not limited to, equine protozoal myelitis, west nile virus infection, and Herpes infection. Sequences have been included to identify the presence of the agents that are caustive of these equine diseases.

[077] Arrays according to the invention can have many uses, including but not limited to: 1. As a research tool to identify changes in gene expression with time, experimental conditions, or other manipulations of any sample, e.g., blood or tissue from horses or cell/tissue cultures, etc., for example, to directly study diseases or conditions of horses or use the horse as a model for human diseases. 2. As a diagnostic test for presence or severity of disease or other conditions, including but not limited to osteoarthritis or other cartilage damage, bone damage, osteochondrosis dessicans, equine protozoal myelitis, equine west nile virus infection, herpes virus infection, etc. 3. As a discovery tool to identify new genes that are involved in certain disease pathways. The features of this test that are superior include the high accuracy, sensitivity and large scale nature of the data input to discriminate among diseases. There is no current alternative for this screening in horses.

[078] All of the compositions and methods disclosed and claimed herein can be made and executed without undue experimentation in light of the present disclosure. While the compositions and methods of this invention have been described in terms of preferred embodiments, it will be apparent to those of skill in the art that variations may be applied to the composition, methods and in the steps or in the sequence of steps of the method described herein without departing from the concept, spirit and scope of the invention. All such similar substitutes and modifications apparent to those skilled in the art are deemed to be within the spirit, scope and concept of the invention as defined by the appended claims.

[079] The following examples are included to demonstrate preferred embodiments of the invention. It should be appreciated by those of skill in the art that the techniques disclosed in the examples that follow represent techniques discovered by the inventor to function well in the practice of the invention, and thus can be considered to constitute detailed modes for its practice. However, those of skill in the art should, in light of the present disclosure, appreciate that many changes can be made in the specific embodiments which are disclosed and still obtain a like or similar result without departing from the spirit and scope of the invention.

[080] **EXAMPLES**

[081] Further details of the invention can be found in the following examples, which further define the scope of the invention.

[082] Example 1 - Construction of Equine Nucleic Acid Database and Microarray

[083] Gene sequences were obtained from the public (GenBank) database, which is maintained at the National Center for Biotechnology Information (NCBI). The sequences were obtained by queries to the GenBank and the returned results were downloaded in GenBank format to the local computer.

[084] The project was completed by using a series of Java application programs which were run under the JAVA™ 2 Runtime Environment, Standard Edition, Version 1.4.1 from the Sun Microsystems, Inc. using a Dell Optiplex GX240 Intel(R) Pentium (R) 4 CPU 1.70 GHz with 256 MB of RAM with Microsoft Windows XP Professional Version 2002 operating system. The BlastN and BlastX were

conducted using the bioinformatics resources at the Ohio Supercomputer Center (http://www.osc.edu). Table 1 lists all the programs used.

Table1. Software and programs used

Name	***Function***
GetEquine	Selects the gene sequences which are from the source of either equus caballus or equus caballus (horse)
CheckCDS	Collects the coding sequences and non-coding sequences separately
GetThreePrimeCompleteCDS	Selects the coding sequences which contain the stop codons at the 3´ ends.
CheckMRNA	Splits the gene sequences into mRNA sequences and DNA sequences
FastaG	Transforms the gene sequences in GenBank format to FASTA format
ClusterG	Identifies the unigene sets; if sequences are found >90% identical match, only the longest sequence is stored
FastaCombine	Combines different FASTA files to one FAST file
GetPolyAEST	Selects ESTs which claim as "PolyA=Yes" or "PolyA=No"
SelectHighQualityEST	Selects the high phred quality region of the ESTs based on the annotated start and stop positions in the GenBank format
GetRC	Obtains the reverse complementary sequence of a target sequence
BlastN	Nucleotide-nucleotide sequence comparison
BlastX	Nucleotide-protein sequence comparison

The source code for each program is provided in Appendix A.

[085] The overall design steps in selecting the 3´ equine annotated genes and ESTs are summarized in Figure 1.

[086] **Construction of equine, and human/mouse sequence databases**

[087] Equine gene sequences were first obtained through a query of "equus caballus" to the GenBank database at the NCBI web site. A total of 20,022 sequences were returned (as of June 2003) and downloaded in GenBank format to the local computer. Program `GetEquine` was performed to specifically select those gene sequences that are from either equus caballus or equus caballus (horse), and 18,924 sequences were obtained and named as "EquusCaballusSequences." This is the original database from which 3′ equine coding sequences and 3′ equine ESTs were identified.

[088] By a query of "homo cds" to the GenBank database at the NCBI web site, 208,480 human sequences (as of the date the Genbank was accessed) were returned and downloaded in GenBank format to the local computer, which were then transformed to FASTA format using the FastaG program. Similarly, by a query of "mus cds," 205,373 mouse sequences (as of the date the GenBank was accessed) were obtained and stored in FASTA format. The resulting human and mouse coding sequences were combined and a correspondent HumanMouseCDS database was created at the Time Logic DeCypher System at the Ohio Supercomputer Center (http://www.osc.edu).

[089] **Selection of 3′ equine coding sequences**

[090] To screen out the 3′ equine cDNA sequences, program CheckCDS was first applied to the EquusCaballusSequences, with 981 equine coding sequences and 17,943 equine non-coding sequences identified, respectively. The equine coding sequences contain both mRNA and DNA sequences. DNA

sequences contain alternative exons and introns, and the latter are removed to produce the mature mRNA. Preferably, mRNA sequences are selected for a gene expression microarray. Program CheckMRNA was performed on the EquusCaballusCDS file, with 436 equine mRNA coding sequences and 545 equine DNA coding sequences identified, respectively.

[091] The equine mRNA coding sequences were further split into two-hundred 5′ partial coding sequences and two-hundred thirty-six 3′ complete coding sequences using the GetThreePrimeCompleteCDS program. 3′ complete coding sequences contain stop codons at the three-prime ends, and hence are either full-length sequences or partial sequences yet 3′ anchored. All these two-hundred thirty-six 3′-anchored sequences were collected for further analysis. Similarly, the equine DNA coding sequences were split into one-hundred thirty-eight 3′ complete coding sequences and four-hundred seven 5′ partial coding sequences. Only the 3′ complete DNA sequences were subjected to further analysis, but 5' DNA partial sequences could be further evaluated if desired. (See Table 2.)

[092] It is quite often that one single gene may be represented by several sequences, each with a different GenBank Accession Number. The same genes may be sequenced and deposited separately by different labs, or the gene sequences may first be deposited into GenBank as partial coding sequences and later as complete sequences. Therefore, multiple sequences, although with different GenBank Accession Numbers, can actually represent the same gene.

[093] To address this potential problem, the FastaG program was first applied to transform the sequences from the GenBank format to the FASTA format,

in which the sequence begins with a single-line description followed by lines of sequence data. Then the ClusterG program was used to identify the unigene clusters and only keep the longest sequence for each cluster. One-hundred ninety-five equine mRNA 3′ complete coding sequence clusters and fifty equine DNA 3′ complete coding sequence clusters were obtained. Because the complete gene (DNA) sequences may contain introns, the virtual respective mRNA sequences of the above equine DNA sequences were obtained by selecting the mRNA or CDS features at the respective GenBank website. The equine mRNA and virtual mRNA sequences were combined with the FastaCombine program and screened again with the ClusterG program for unigene clusters and the final 209 equine annotated 3′ coding sequences were identified. These equine sequences are either full-length sequences or 3′ anchored.

[094] This screening was based on selecting the 3′-biased coding sequences. However, some partial sequences may actually contain regions close to the 3′ end and thus could also be suitable for use in a microarray. To capture these sequences, the two-hundred 5′ partial equine mRNA coding sequences were first reduced to 149 clusters with the ClusterG program. Sequence comparisons of these clusters were performed against the HumanMouseCDS database using the BlastN program at the Time Logic DeCypher System at Ohio SuperComputer Center. The blast result was manually examined and a total of 83 equine partial coding sequences which are in close proximity (i.e., within 500 bp) to the 3′ end or important to our research were identified and combined with the previously identified 209 3′ equine coding sequences using the FastaCombine program. Program

ClusterG was performed on the combined sequences and 311 final equine annotated gene sequences were selected for the microarray. Table 2 summarizes the result in each step of selecting the 3′ equine coding sequences.

Table 2. Summary in identifying the 3′ equine coding sequences

Sequence	*Number*
Equine sequences	18,924
Equine coding sequences	981
Equine coding sequences, mRNA	436
Equine coding sequences, mRNA, 3′ complete	236
Equine coding sequences, mRNA, 3′ complete cluster	195
Equine coding sequences, mRNA, partial	200
Equine coding sequences, partial mRNA selected	83
Equine coding sequences, DNA	545
Equine coding sequences, DNA, 3′ complete	138
Equine coding sequences, DNA, 3′ complete cluster	50
Equine coding sequences, selected mRNA and DNA	328
Equine coding sequences selected	311

[095] The selected annotated equine gene sequences were also subjected to the BlastX assay against the SwissProt database (Gasteiger E, Jung E, Bairoch A, "SWISS-PROT: connecting biomolecular knowledge via a protein database," *Curr Issues Mol Biol* 3(3):47-55(2001)) to confirm the sequence orientation, and all sequences were shown in the sense orientation (data not shown).

[096] **Selection of 3′ equine ESTs**

[097] The 3′ equine ESTs were isolated from the 17,943 equine non-coding sequences. Candidate 3′ equine ESTs were first obtained using the GetPolyAEST program against the EquusCaballusSequences. Program GetPolyAEST selects the EST sequences which indicate as "PolyA=Yes" or "PolyA=No". As noted above, the sequence information from these ESTs may contain the polyA tail if the sequencing process reaches to the 3′ end. However, if the sequencing is initiated at the 5′ end

and stops in the middle, the obtained sequence information may not include the polyA tail, although it may be very close to the 3′ end. Therefore, ESTs claiming "PolyA=No" may not necessarily mean that they are not at or close to the 3′ end. Based on this, we first selected the ESTs which claim both "PolyA=Yes" and "PolyA=No" so that a maximal pool of candidate 3′ ESTs could be constructed. A total of 8,752 putative equine 3′ ESTs were obtained. Then the SelectHighQualityEST program was applied to specifically select the high-quality, vector-trimmed regions and transform into FASTA format.

The resulting high quality ESTs (8,752 sequences) were subjected to the ClusterG program to obtain EST clusters (4,139 clusters). Table 3 shows the 3′ ESTs. (We selected the longest sequence for each cluster. Longer sequences can be obtained by sequence assembly. For long sequences, the whole sequence is fragmented and each fragment is sequenced individually and the whole sequence is obtained by assembly later. Some sequencing may be performed in both directions. Through assembly, more complete sequences can be obtained, if there is enough overlap exists between the fragments.)

Table 3. Summary of identifying equine 3′ ESTs

Sequence	*Number*
Equine sequences	18924
Equine EST with PolyA=Yes or PolyA=No	8752
Equine PolyA (Yes/No) EST cluster	4139
Equine EST cluster with BlastN hit	3791
Equine EST screened	3155
Sense EST	2856
Antisense EST	299
Equine coding sequences selected	311
Final equine sequences selected for the array	3288

[098] To obtain the annotations and 3′ bias confirmation, the equine ESTs were blasted against the HumanMouseCDS database using the BlastN algorithm at the Ohio SuperComputer Center facility. A total of 3,791 equine EST clusters had blast hits, with 3046 having blast score higher than 60. Of these, only sequences with blastE values of $<10^{-8}$ were considered candidates for selection. (Makabe KW, Kawashima S, Minokawa T, Adachi A, Kawamura H, Ishikawa H, Yasuda R, Yamamoto H, and Kondoh K, et al. "Large-scale cDNA analysis of the maternal genetic information in the egg of *Halocynthia roretzi* for a gene expression catalog of ascidian development, *Development* 128:2555-2567 (2001)) . The blast result also was examined manually to remove any ESTs that matched to the 5′ end of the corresponding human or mouse coding sequences. Among the other 1093 clusters with low or no blast scores, 441 have either polyA tail or polyA signals. These sequences were combined with the above selected ESTs with higher blast scores and subjected to cluster analysis. A total of 3,155 ESTs were identified as 3′ biased. The orientations of the ESTs were also derived from the blast results by inspection of the direction of the sequence match (blast hit), with 2,856 in sense orientation and

299 in antisense orientation (Table 3). The reverse complementary sequences of the antisense ESTs were obtained by the program GetRC and were combined with the sense equine ESTs. The resulting ESTs were also combined with the annotated equine coding sequences and undergone the cluster analysis again. Total of 3,288 equine 3′ coding sequences and 3′ ESTs were selected for the equine gene expression microarray.

[099] Note that many of the annotated genes that were publicly available were from laboratories studying musculoskeletal conditions. In total, this may include 100-200 genes. Thus, in the end, the collection of sequences had a slight bias toward musculoskeletal genes.

[0100] Table 5 lists an annotation for the equine sequences identified in accordance with the invention; Table 6 shows the equine sequences.

[0101] **Preparation of the Microarray**

[0102] The probe set design was accomplished based on the selected equine sequences according to Affymetrix's chip design guide. The probe sets were selected by the following parameters: probe set score, gap multiplier, cross hybridization multiplier, probe count, raw standard deviation, siflength, etc. Each sequence was checked for unique, identical, or mixed probe sets. Probe sets with a score no less than 2.0 for unique set or a score no less than 4.0 for identical or mixed set were selected. A total of 68,266 equine oligonucleotide probes were included on a high density microarray, with average 11 perfect matches and 11 single nucleotide mismatches for each equine gene.

[0103] **Discussion**

[0104] Genetic information has been exploding dramatically since its construction. At the time of the equine microarray design, over 20,000 equine sequences were available in the public database (GenBank). How to data-mine the 3′-biased sequences is an issue in generating gene expression microarrays, including equine microarrays. Here, we have disclosed a unique computer-based approach that is applicable for creating gene expression microarrays for any other species. The approach generally involves two major steps: identifying the 3′ coding sequences and 3′ ESTs.

[0105] In identifying the 3′ equine coding sequences, we first focused on the selection of full-length coding sequences and partial sequences with 3′ end. This is done by selecting the coding sequences with the stop codon at the 3′ end. This approach ensures that sequences selected are 3′ anchored. Some of them also contain the 3′-untranslated regions, which may be more species-specific compared to the coding region. To capture additional coding sequences for the microarray, we performed the blast analysis for the partial coding sequences against the self-constructed HumanMouseCDS database instead of the non-redundant (nr) nucleotide database available at NCBI. The HumanMouseCDS database is actually a subset of the nr database. Most of the sequences are annotated human or mouse coding sequences. Therefore, the blast result based on this database provides more useful information, which was especially valuable in the equine EST annotation and sequence orientation determination. Moreover, as the HumanMouseCDS

database is much smaller, the computing time for the blast assay is tremendously decreased.

[0106] One approach in constructing the cDNA library used for transcript sequencing is using the oligo-dT as the primer in the first strand cDNA synthesis. This would preferentially begin sequencing from the 3′ end due to priming on the poly A tail. In other methods, the sequence information from these ESTs may contain the polyA tail if the sequencing process reaches to the 3′ end. However, if the sequencing is initiated at the 5′ end and stops in the middle, the obtained sequence information may not include the polyA tail, although it may be very close to the 3′ end. Therefore, ESTs claiming "PolyA=No" may not necessarily mean that they are not at or close to the 3′ end. Based on this, we first selected the ESTs which claim both "PolyA=Yes" and "PolyA=No" so that a maximal pool of candidate 3′ ESTs could be constructed.

[0107] ESTs are short sequences, representing only fragments of genes, not complete coding sequences. The sequences may be in either sense or antisense orientation. Therefore, a major effort and emphasis is focused on how to best annotate these ESTs. In fact, we first annotated the equine ESTs with blast analysis against the nr database (data not shown). However, an overwhelming number of hits occurred between the ESTs and sequences without much useful information, as the hits occurred with the chromosomal sequences, cDNA clones, etc. Therefore, we modified the blast analysis against the self-constructed HumanMouseCDS database that contained more concentrated annotated human and mouse coding sequences. Approximately 92% of the ESTs had blast hits and putative annotations

were provided. (Annotations were categorized based on the published papers. Escribano J. and Coca-Prados, M., "Bioinformatics and reanalysis of subtracted expressed sequence tags from the human ciliary body: Identification of novel biological functions," *Molecular Vision* 8:315-332 (2002); Lo J. et al. "15,000 Unique Zebrafish EST Clusters and Their Future Use in Microarray for Profiling Gene Expression Patterns During Embryogenesis," *Genome Research* 13(3):455-466 (2003)). (See Table 4.)

[0108] For the gene expression microarray, further probe design could be based on the antisense strand of the selected sequences. The array can be either cDNA spotted microarray (the clones can be purchased or self obtained by PCR) or the Affymetrix oligonucleotide GeneChip. cDNA spotted microarrays use longer sequences as probes which are advantageous in that sequences could be spotted first without being known and the gene sequence of interest could be determined later. However, this approach is labor intensive and costly in producing and maintaining the clones or PCR products. Errors may occur in mis-assigning the clones. (Halgren RG, Fielden MR, Fong CJ, and Zaxharewski, TR, "Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones," *Nucleic Acids Research* 29:582-588 (2001).)

[0109] It is difficult to distinguish closely related gene families using cDNA microarray. Also, for rarely expressed genes, it is hard to obtain the suitable cDNA clones. On the other hand, if the sequence information is available, oligonucleotides can be synthesized to hybridize specifically and uniquely to any available target genes. This approach avoids the need to manipulate large cDNA clone libraries.

The cross-hybridization problem due to the short length of the probe could be ameliorated by the usage of several probe sets per gene. In the Affymetrix GeneChip system, the use of perfect match and mismatch design provides a control for background noise and cross-hybridization from unrelated targets. The chip cost has now decreased several-fold and become more affordable to academics, compared to large-scale cDNA microarrays.

[0110] This is the first published microarray accumulation of equine annotated genes and ESTs and all that is publicly available to date. The equine chip includes equine gene sequences functioning in apoptosis, cell cycle, signal transduction, developmental biology, etc, as listed in Table 4. (Escribano J. and Coca-Prados, M., "Bioinformatics and reanalysis of subtracted expressed sequence tags from the human ciliary body: Identification of novel biological functions," *Molecular Vision* 8:315-332 (2002); Lo J. et al. "15,000 Unique Zebrafish EST Clusters and Their Future Use in Microarray for Profiling Gene Expression Patterns During Embryogenesis," *Genome Research* 13(3):455-466 (2003)).

Table 4. Category of the selected equine 3′ coding sequences and 3′ ESTs

Protein category	*3′ coding sequence*	*3′ EST*
Enzyme		
Dehydrogenase	4	35
Isomerase		15
Kinase	1	78
Phosphatase		39
Synthase	5	35
Transferase	1	37
Oxidase		8
Peptidase		6
Others	14	69
Protein Synthesis		
Ribosomal protein	5	105
Initiation, elongation,		49

Protein category	*3´ coding sequence*	*3´ EST*
and other factors		
RNA binding	2	108
DNA binding	5	203
Transcription factor	3	64
Protein degradation	8	62
Membrane protein	2	53
Cellular Signaling		
Receptor and related	32	223
Ligand and other exchange factors	62	142
Structural protein	21	98
Cell division	5	41
Cell adhesion	2	12
Cell differentiation	2	15
Ligand binding or carrier	5	102
Transporter	8	74
Antioxidant	1	6
Immune-related proteins	33	152
Lipoprotein	1	3
Apoptosis	1	24
Chaperone	3	21
Enzyme inhibitor	7	33
Enzyme activator		10
Developmental protein	4	27
Motor		10
Unclassified	30	1038

[0111] Data from this microarray will provide insight into gene expression for equine specific diseases and conditions. Thousands of equine ESTs whose genetic functions were unknown previously are now annotated. This not only enriches the equine gene expression profile, but also will provide a solid base for future full-length gene discovery and analysis.

[0112] Example 2 - Construction of Canine Nucleic Acid Database and Microarray

[0113] Example 1 was repeated to create a canine database, with some alterations made in the procedure. First, due to the limitation of the microarray size and the very large number of canine sequences publicly available, only the fully annotated 3'-complete mRNA canine coding sequences were selected. No canine ESTs were included in the processing. Otherwise, the steps were similar as in Example 1: GetCanine --> GetCDS --> CheckMRNA --> GetThreePrimeCompleteCDS --> FastaG --> ClusterG.

[0114] Example 2 - Equine-Specific Large-Scale Gene Expression Analysis of Developmental Bone Diseases

[0115] Developmental Orthopedic Disease (DOD) represents a group of bone diseases that manifest during growth and development and include articular dyschondroplasia (osteochondrosis dessicans, OCD) and cervical vertebral malformation (CVM). The underlying pathogenesis is altered endochondral ossification of mineralizing cartilage. Site-specific clinical syndromes result. Abnormalities at the articular growth front result in a dyschondroplasia called osteochondrosis dessicans (OCD) or intra-articular cartilage flaps with abnormal underlying bone. The incidence of articular osteochondrosis is increasing and the condition is present in the horse population at high levels (10-25%). OCD induces arthritis and lameness and is usually treated surgically. The hock and stifle are the most common joints affected. Abnormalities of vertebral growth result in narrowing of the cervical vertebral canal in combination with malformation of the vertebra. The result is spinal neurologic disease characterized by ataxia and weakness.

[0116] The syndrome is termed cervical vertebral stenotic myelopathy (CVM) and is treated with anti-inflammatory medication, nutritional support, and, in selected cases, surgical cervical fusion. CVM is the leading cause of noninfectious spinal cord ataxia in the horse and affects 2% of the Thoroughbred population. Both conditions are distributed internationally, in multiple breeds and usually manifest in the young growing horse. Studies supporting a genetic predisposition to both conditions, and unique biochemical and molecular features of osteochondrotic cartilage in horses, suggest that evaluation of gene expression will be a productive approach to identifying the presence and predisposition to this disease. The use of microarrays for gene expression studies and diagnostics is becoming well established. The use of a species-specific microarray is of critical importance for accurate biomarker identification and monitoring of highly specific markers. In cross-species hybridization on microarrays, even single nucleotide mismatches can alter the detectable gene expression and relative intensities resulting in erroneous conclusions. Affymetrix is a recognized manufacturer of large-scale microarray technology that is sensitive, specific, and highly repeatable.

[0117] We have collected preliminary data on a preeminent Kentucky thoroughbred farm (> 100 foals/year) in collaboration with the farm veterinarians. Thirteen yearlings with OCD and 7 age- and sex-matched yearlings (within a month of age); and 6 weanlings with CVM, 3 weanlings with CVM affected siblings, and 4 age and sex match control weanlings have yielded high quality RNA from blood (O.D. 260/280 >2.0) that is frozen at –80C. The investigators have collected all clinical data including radiographs, myelograms, lameness, and neurologic

examinations. Gene expression analysis using this equine-specific microarray is underway.

[0118] This example describes how to quantify and bioinformaticly analyze gene expression alterations associated with two of the most common developmental orthopedic diseases in young horses, articular dyschondroplasia (osteochondrosis dessicans, or OCD) and cervical vertebral malformation (CVM). Gene expression markers will be identified that can uniquely identify the presence of these disease conditions (a signature). This example describes the construction of a bioinformatic tool that can predict, diagnose, and monitor therapy of these conditions. The hypotheses are that a unique gene expression profile will exist for OCD and CVM during growth and development and will consist of the up or down regulation of tens to hundreds of genes, particularly genes in the skeletal regulation pathway.

[0119] First, gene expression is bioinformatically profiled to identify a gene expression signature for OCD and CVM for use as a diagnostic tool. Second, gene expression is analyzed to predict clinical disease in a prospective multiyear cooperative study for use in disease prevention and monitoring.

[0120] In detail, blood (n=20), synovial fluid (for OCD, n=10), and cerebral spinal fluid (for CVM, n=10) from 20 affected weanling or yearling horses, as well as 20 age-matched normal control horses, is processed. Cells are evaluated from blood, and synovial fluid (for OCD) or cerebral spinal fluid (for CVM) from horses with clinically and radiographically apparent OCD or CVM (including myelographic evidence of stenosis) compared to unaffected matched control horses. Hock or stifle joints with effusion is accepted for OCD and neurologic scores $\geq$2 for CVM.

[0121] Blood from 30 Thoroughbred foals, 30 weanlings, and 30 yearlings over two-years, on the same large breeding farm in which all weanlings and yearlings are screened for OCD and CVM and the prevalence of these conditions is known, are processed. Year-1: 30 foals and again as weanlings; Year-2: The same 30 as yearlings. Selection bias based on historical records enhance the chance to obtain ~ 50% that will develop OCD or CVM. Once clinical disease is diagnosed, animals are assigned as affected and the blood from the closest unaffected age and sex matched control is identified.

[0122] Cells from these fluid samples are isolated by centrifugation (SF and CSF) or manual buffy coat fractionation from blood (Sequire Kit, PPAI Medical, FL) and batch processed for ribonucleic acid (RNA) extraction (Qiagen RNAeasy), cDNA synthesis, in vitro transcription, RNA amplification and fragmentation, and RNA fluorolabeling as per the GeneChip Expression Analysis Technical Manual, Affymetrix, Inc., 2001. All equipment (Affymetrix hybridization chamber, fluidics station, and computer workstation and software) are publicly available.

[0123] Labeled RNA are hybridized to equine species-specific high density DNA probes and scanned for gene expression intensity using an Affymetrix Gene Expression System and the equine custom microarray described in Example 1. This equine gene expression microarray represents 3,288 annotated equine genes that contain a bias for musculoskeletal relevance. Over 360 genes represent cell signaling functions, 322 are enzymes, 154 in protein synthesis, 375 in RNA/DNA binding including transcription factors, 193 in cell differentiation including developmental protein function, and 24 in apoptosis pathways. All known relevant

genes to OCD in horses, such as PTHrP, Indian hedgehog, bone morphogenetic proteins, and receptor-activated nuclear factor Kβ ligand (RANK L) are on the array.

[0124] Bioinformatic analysis of gene intensity data by cluster analysis and comparisons among groups (OCD vs control; CVM vs control; OCD vs CVM) is performed using MicroArray Suite 5.0 and Data Mining Tool software (Affymetrix, Inc) for each gene determined to have increased or decreased expression across time (Experiment B) or among groups (Experiments A and B).

[0125] Example 3 - Canine Microarray Gene Expression Analysis for Molecular Therapy of Hip Disease

[0126] Dogs are human's best friends. There are about 300 different dog breeds in the world as a result of a long history of gene pool selection and mixing. The modern domestic dog is unique for the study of human genetic diseases in that it has a larger pedigree than that of the small, outbred human families. Moreover, many of the ~360 known canine genetic diseases are homologs of the human disorders, including osteoarthritis secondary to hip dysplasia. These genetically complicated disorders are not fully controlled by a single gene and are suited for large-scale gene expression profiling to gain insight into the cross-talk associated with the abnormal phenotype. The use of microarrays for gene expression studies and diagnostics is becoming well established. The use of a species-specific microarray is of critical importance for accurate biomarker identification and monitoring of highly specific markers. In cross-species hybridization on microarrays, even single nucleotide mismatches can alter the detectable gene expression and relative intensities resulting in erroneous conclusions.

[0127] Canine disease gene cloning and characterization is the major limiting step in understanding the canine diseases at the gene level. In our preliminary analyses, the current public nucleotide database (GenBank) has stored close to 2 million canine related genetic records, while only 0.1% have been annotated with genetic function. Most nucleotide entries are unknown chromosomal sequences and expressed sequence tags of unknown function. With the maturation of primarily the human and mouse databases, tens of thousands of gene sequences have been functionally identified and mapped. Such information can be used to decipher the canine sequences through comparative analysis.

[0128] In this example, we describe the design and use of a canine database, similar to that described for equine in Example 1. The design annotates sequences by Blast to a human/mouse coding sequence database, trims for high quality sequence, substantially reduces duplication, and selects for 3′ complete sequencing to permit high resolution probe design critical for ribonucleic acid (RNA) detection by current technology that involves 3′ amplification.

[0129] Osteoarthritis (OA) is a debilitating disease affecting both canine and human patients. It is one of the most common sources of chronic pain treated by veterinarians, estimated to affect one in five of 68 million adult dogs and commonly affects the hip joint secondary to hip dysplasia. Accordingly, the incidence of musculoskeletal pathology in dogs less than one year of age has been estimated at 22%, often related to hip dysplasia. Use of large-scale gene expression profiling of osteoarthritic cartilage to assess phenotype and alterations with experimental manipulation are beginning to appear in the literature, including IL-1.

[0130] This example describes the generation of an exhaustive canine database for gene expression and applies this information to large-scale microarray analysis to assess the ability of molecular therapy to promote a regenerative phenotype in canine osteoarthritic (OA) cartilage. This example captures current state of the art technology made possible from the recent canine genome sequencing projects for both public academic use and the use in profiling inducible cellular dedifferentiation pathways of OA chondrocytes.

[0131] The current > 1.5 million canine sequences on the public database will likely condense to < 40,000 high quality, unique annotated canine sequences most of which will contain the criteria necessary for inclusion on a microarray, such as 3´-bias, and also, the bone morphogenetic protein-2 (BMP-2) in combination with interleukin-1 receptor antagonist (IL-1ra) will induce gene expression patterns involving hundreds of genes that profile a healthier chondrocyte phenotype, including aggrecan and type II collagen up-regulation and metalloproteinase down-regulation.

[0132] This example describes the curation, pruning, and annotation of the public canine nucleotide database so it can be used for further canine genomic functional analysis or for generating canine species-specific large-scale gene expression microarrays. These data may complement imminent commercial canine high-density microarrays, and allow for comparison of gene expression patterns of OA hip cartilage from dysplasic dogs that have been genetically engineered to express BMP-2 and/or IL-1ra as a measure of an induced de-differentiation gene expression profile typical of more healthy chondrocytes. This example proves initial

efficacy of novel molecular therapies for hip dysplasia that can be delivered by joint injection, offering a pain-relieving and disease-modifying therapy.

[0133] The approach used to obtain the equine database was through queries to NCBI, and downloaded the result to the desktop computer. For equine sequences (~20,000 records), this is acceptable. However, for dog, it may be difficult to download ~ 2 million records in GenBank format from the web to the local computer (PC) by query. Thus, for canine genomic sequences, a file transfer protocol can be used instead to directly transfer the file from NCBI.

[0134] In detail, a canine nucleotide sequence database is obtained from GenBank through file transfer protocol (ftp://ftp.ncbi.nih.gov). As described in Example 1, Java-based software programs are used to sequentially: 1) curate sequences specific to canis familiaris, 2)select coding sequences, 3)select high-quality, vector-trimmed regions of expressed sequence tags (ESTs), 4)convert to FASTA format, 5) prune by cluster analysis to eliminate duplication, and 6)select sequences with complete 3′ sequencing. For annotation and sense orientation confirmation, the canine ESTs are blasted against a similarly generated Human/MouseCDS using the BlastN algorithm at the Ohio SuperComputer Center facility. Sequences below the threshold E value ($< 10^{-8}$) are selected for further annotation. Annotated sequences are blasted against the fully annotated SwissProt protein database to further confirm annotation and sequence orientation. Table 7 lists an annotation of the canine sequences identified in accordance with the invention; Table 8 shows the canine sequences.

[0135] Next, minced cartilage explants are developed from canine hips with OA removed at total joint replacement and cartilage from normal hips of similar aged dogs with no evidence of OA. Explants are divided into 5 groups and cultured for 14 days: 1)untreated, 2) adenoviral control treated; cultured with a nonreplicative adenoviral (Ad) vector containing a noninductive marker gene for beta-galactosidase (Ad-LacZ), 3) treated with Ad-human (h) BMP-2;a cartilage morphogen, 4) treated with Ad-hIL-1ra; blocks cartilage degradation, and 5) treated with both Ad-hBMP-2 and Ad-hIL-1ra. Outcome assessments include gene expression analysis by microarray using the inventive canine array and/or a custom array generated by oligosynthesis of sequences known to be of importance in OA and chondrogenesis. Explants are then histochemically stained for assessment of morphology and aggrecan and type II collagen production.

[0136] Example 4 - Selection of Viral and Protozoal Sequences for Inclusion on Microarray

[0137] Equine viral and protozoal diseases were identified for use in a diagnostic microarray. The selected organisms included equine herpesvirus 1, equine herpesvirus 2, equine herpesvirus 3, equine herpesvirus 4, equine herpesvirus 5, equine morbillivirus, Neospora hughesi, Sarcocystis neurona, and West Nile virus. Nucleic acid sequences were selected based on the following procedure.

[0138] Briefly, the herpesviruses 1, 2, and 4, and West Nile had complete genome data available in the public database. Therefore, for these, the sequences encoding capsid, membrane, envelope, or virus package proteins were specifically

selected. Other viruses did not have complete genome data, so all of the available sequences were selected for those species. Table 9 lists an annotation of equine viral and protozoal sequences identified in accordance with the invention; Table 10 shows the actual sequences.

[0139] The sequences can be used as is, as the basis for a microarray, or can be separated based on pathogen and then used for generation of a microarray.

[0140] Example 5 – Equine-specific Large Scale Gene Expression Analysis of Equine Protozoal Myelitis

[0141] Equine protozoal myelitis represents an infectious disease with protozoan organisms, sarcocystis neurona, canis neospora, and maybe others, that encyst in neuronal cell bodies in the central nervous system resulting in neurologic disorders in horses. The horse is a dead-end host and probably not a host in the primary life cycle of the organisms. Well-described clinical signs include spinal ataxia and weakness as well as muscle atrophy, peripheral nerve dysfunction, and possibly any other lower motor neuron dysfunction. Diagnosis is usually inconclusive and limited because organisms are hard to find on histology due to lesion rarity in the CNS and obviously requires death of the animal to retrieve the brain and spinal cord. Blood and cerebral spinal fluid assays to date are inconclusive because they have depended on antibody titers or staining that does not effectively distinguish exposure to organisms and pathologic invasion by the organism. Other diagnostic approaches to identify organisms have been limited by oversensitivity (high false positives) and failure to assess the biologic response to the organism as part of the cause of the development and severity of the disease.

[0142] The use of a species-specific large-scale gene expression microarray permits the simultaneous measurement of the biologic response to the organisms, which may include increased inflammatory and immunologic responses. Cells from spinal cord fluid or blood are processed for use on the array to identify these changes and monitor response to treatment. RNA placed on the microarray provides a signature gene expression typical of the disease as compared to other neurologic diseases such as CVM previously described.

[0143] Example 6 – Protozoan-specific Gene Expression Microarray Analysis for Equine Protozoal Myelitis

[0144] Sequences have been placed on the array, which are genes expressed by Sarcocystis neurona and Canis neospora, similarly obtained from the public database as the sequences in Example 4 above. These S. neurona and C. neospora RNA sequences are selected to identify as high sensitivity as can be obtained on a microarray the presence of the organism and its infection in cells of the horse or other species for that matter. Since these sequences are generated by the organisms, the species from which infected tissue was obtained would not be required to be only horse.

[0145] The equine species has a significant prevalence of this disease and therefore would be a logical animal to inspect tissues. The sequences on the microarray are specific to these organisms. In the case of viruses, the organisms must have infected cells to make this RNA that would be detected on the array.

[0146] Other diagnostic tests for the presence of these organisms have attempted to detect DNA from the organisms, by PCR or other techniques. DNA is

highly stable and can represent dead or silently encysted organisms. DNA-based techniques are also known for a high false positive rate due their extreme sensitivity and ease of laboratory or processing contamination. RNA, on the other hand, is labile and to be present, must be from active organisms. It is not normally a contaminant laboratories, as it is readily degraded at room temperatures.

[0147] Example 7 – Equine Viral-specific Gene Expression Analysis of Herpes Virus –1 Infection in Horses

[0148] Equine Herpes Infection is classically characterized by fever, nasal discharge (i.e., an upper respiratory tract infection) and malaise. This disease, however, can be particularly virulent with some strains, such as occurred in 2003 at Findlay College Equestrian Program herd in Central Ohio. The Ohio State University was integrally involved in containment of this outbreak and in the diagnostics.

[0149] Of 132 horses, the majority developed clinical signs (>~75%), and this is an exceptionally high virulence rate. Typically, most exposed horses will not develop clinical signs, but fight off the invading organism before clinical signs occur. Of these, a high percent (>10%) developed the complicating neurologic disease that is associated with this virus, documenting it as a neurotrophic strain. Diagnosis is currently dependent on serum antibody titer and viral culture from nasal swabs. The former is limited by representing past exposure only, not current disease. Therefore, serial titers are necessary to demonstrate expected increases in titers.

[0150] In all regards, these results can be influenced by previous vaccination status as most horses are vaccinated for equine herpes–1. The viral culture

requires a minimum of 2 weeks and typically longer to complete. It is fraught with false positives from organisms harboring in the laboratory and contaminating long-standing culture plates. This was a problem in the diagnostic testing of this outbreak. Use of Herpes virus-1 RNA sequences on a microarray for testing offers increased sensitivity, bulk analysis, and rapid turnaround.

[0151] RNA from cells from nasal swabs and mucus or any other tissue suspected of containing organisms, such as spinal cord, cerebral spinal fluid cells, blood, discharges, etc. is isolated and placed on the microarray of Example 4, with appropriate control samples. The presence of herpes virus-1 RNA means that the organism is not only present but has infected cells, inserted its DNA into the cell nucleus and is using the cell machinery to make the virus's own RNA to make the virus's own proteins necessary for it to invade and replicate. In other words, the virus has infected the host, and is not just present. It currently takes 3 days to complete the processing for this microarray and obtain results, a substantial savings in time as compared to several weeks. The same tests can be run on equine morbillivirus, Neospora hughesi, Sarcocystis neurona, and West Nile virus.

[0152] This microarray diagnostic test also can detect infection before clinical signs even become apparent. The importance of early diagnosis includes rapid isolation of infected animals, release of uninfected animals from expensive quarantine, identification of outbreaks, and moving animals at high risk for the complications like neurologic disease and abortion.

[0153] Other embodiments of the invention will be apparent to those skilled in the art from consideration of the specification and practice of the invention disclosed

herein. It is intended that the specification and examples be considered as exemplary only, with a true scope and spirit of the invention being indicated by the following claims.

WHAT IS CLAIMED IS:

1. A method of preparing a species-specific nucleic acid database comprising:

- selecting from a species-non-specific nucleic acid database species-specific nucleic acids comprising coding sequences;

- selecting from a species-non-specific nucleic acid database species-specific nucleic acids comprising noncoding sequences;

- selecting from the coding sequences those sequences that are 3′-compete or 3′-coding biased, wherein 3′-coding biased sequences comprise 5′-partial sequences having desirable characteristics;

- selecting from the noncoding sequences those sequences that include poly-A tails or are derived from sequences that include poly-A tails;

- reducing redundancy in selected sequences;

- comparing sequences comprising unannotated sequences to a collection of sequences comprising annotated coding sequences and selecting those sequences satisfying a threshold of similarity; and

- collecting all selected sequences.

2. The method according to claim 1, wherein the species-specific nucleic acid database is an equine-specific nucleic acid database.

3. The method according to claim 1, wherein the species-non-specific nucleic acid database is GenBank.

4. An array comprising a plurality of oligonucleotide probes designed to be complementary to and hybridize under stringent conditions with a gene listed in at least one of Tables 5, 7, or 9.

5. An array comprising a plurality of oligonucleotides, wherein:

a) the oligonucleotides are chosen from the nucleic acid sequences shown in Tables 6, 8, or 10, and wherein the array comprises 10 or more of said oligonucleotides; or

b) the oligonucleotides comprise nucleotide probes designed to be complementary to, or hybridize under stringent conditions with, 10 or more nucleic acid sequences shown in Tables 6, 8, or 10.

6. The array according to claim 5, wherein the oligonucleotides comprise nucleotide probes designed to be complementary to, or hybridize under stringent conditions with, 1000 or more nucleic acid sequences shown in Table 6.

7. The array according to claim 6, wherein the oligonucleotides comprise nucleotide probes designed to be complementary to, or hybridize under stringent conditions with, 2000 or more nucleic acid sequences shown in Table 6.

8. The array according to claim 7, wherein the oligonucleotides comprise nucleotide probes designed to be complementary to, or hybridize under stringent conditions with, 3000 or more nucleic acid sequences shown in Table 6.

9. A method for populating a database of species-specific nucleic acid sequences, comprising

querying a database of nucleic acid sequences to identify nucleic acid sequences associated with a subject species;

processing the identified sequences to create a first subset containing coding sequences and a second subset containing non-coding sequences;

dividing the first subset into a plurality of DNA sequences, if present, and a plurality of mRNA sequences;

processing the plurality of DNA sequences to derive a plurality of virtual mRNA sequences;

dividing the plurality of mRNA sequences into a plurality of complete and mRNA 3' partial sequences, and a plurality of mRNA 5' partial sequences;

processing the plurality of mRNA 5' partial sequences to identify a subset of mRNA 5' partial sequences, each member of the subset satisfying a threshold level of completeness;

identifying members of the second subset containing non-coding sequences that correlate with at least one known coding sequence of at least one species other than the subject species; and

combining the plurality of virtual mRNA sequences, the plurality of complete and MRNA 3' partial sequences, the subset of mRNA 5' partial sequences, and the identified correlated sequences to create the database of species-specific nucleic acid sequences.

10. The method according to claim 9, wherein the step of identifying includes comparing each member of the second subset to each member of a database containing annotated human nucleic acid sequences.

11. The method according to claim 9, wherein the step of identifying includes comparing each member of the second subset to each member of a database containing annotated human and mouse nucleic acid sequences.

12. The method according to claim 11, wherein the database containing annotated human and mouse nucleic acid sequences is derived from the database of nucleic acid sequences.

13. The method according to claim 9, further comprising eliminating duplicates within the database of species-specific nucleic acid sequences.

14. The method according to claim 9, further comprising populating the database of species-specific nucleic acid sequences with selected species-specific virus definitions.

15. The method of claim 9, further comprising verifying that each of the identified correlated sequences is represented in sense format.

16. The array according to claim 5, wherein the array is designed for diagnosis of disease.

17. The array according to claim 16, wherein the array is designed for diagnosis equine or canine disease.

18. The array according to claim 5, wherein the array comprises at least one gene or sequence shown in Table 9 or 10, and wherein the array is designed for diagnosis of disease in any tissue of any animal.

ABSTRACT OF THE DISCLOSURE

[0154] Methods of preparing biological databases, and databases prepared according to those methods. In some embodiments, the methods can be performed entirely using computer resources, relying solely on publicly available biological sequence information. The methods of the invention can be used to generate species-specific nucleic acid microarrays.

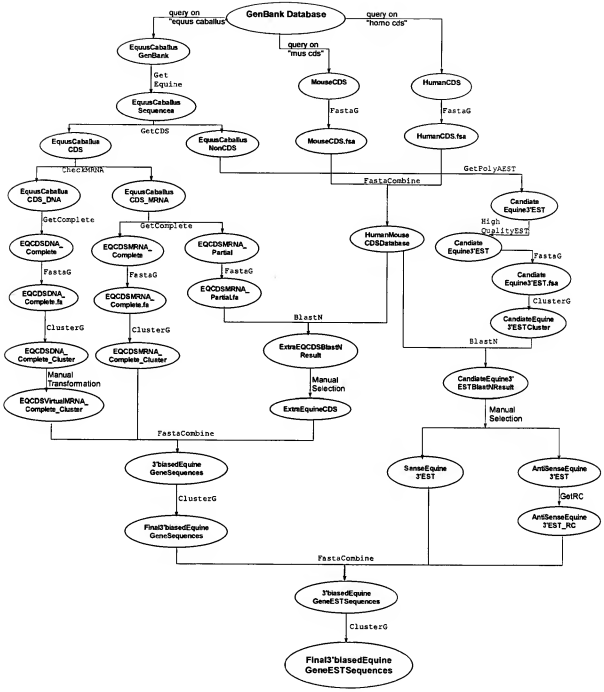

GenBank Database
query on "equus caballus"
query on "mus cds"
query on "homo cds"
EquusCaballus GenBank
Get Equine
EquusCaballus Sequences
MouseCDS
HumanCDS
FastaG
FastaG
GetCDS
EquusCaballus CDS
EquusCaballus NonCDS
MouseCDS.fsa
HumanCDS.fsa
CheckMRNA
FastaCombine
GetPolyAEST
EquusCaballus CDS_DNA
EquusCaballus CDS_MRNA
CandiateEquine3'EST
GetComplete
GetComplete
HumanMouse CDSDatabase
High QualityEST
EQCDSDNA_ Complete
EQCDSMRNA_ Complete
EQCDSMRNA_ Partial
Candiate Equine3'EST
FastaG
FastaG
FastaG
CandiateEquine3'EST.fsa
EQCDSDNA_ Complete.fa
EQCDSMRNA_ Complete.fa
EQCDSMRNA_ Partial.fa
ClusterG
BlastN
CandiateEquine 3'ESTCluster
ClusterG
ClusterG
BlastN
EQCDSDNA_ Complete_Cluster
EQCDSMRNA_ Complete_Cluster
ExtraEQCDSBlastN Result
CandiateEquine3' ESTBlastNResult
Manual Transformation
Manual Selection
Manual Selection
EQCDSVirtualMRNA_ Complete_Cluster
ExtraEquineCDS
FastaCombine
3'biasedEquine GeneSequences
SenseEquine 3'EST
AntiSenseEquine 3'EST
ClusterG
GetRC
Final3'biasedEquine GeneSequences
AntiSenseEquine 3'EST_RC
FastaCombine
3'biasedEquine GeneESTSequences
ClusterG
Final3'biasedEquine GeneESTSequences

Figure1